

This listing of claims will replace all prior versions, and listings, of claims in the application:

1 Claim 1 (currently amended): A system for providing
2 capitalization correction for unstructured excerpts,
3 comprising:
4 a preprocessor to tokenize an excerpt of
5 unstructured content into a set of words; and
6 a capitalizer to analyze the set of words for
7 correct capitalization, comprising:
8 an evaluator to evaluate individual characters
9 constituting at least one such word in the set of
10 words; and
11 a filter to skip the at least one such word if
12 determined to be of a predefined type such that the
13 capitalizer does not capitalize the at least one such
14 word.

1 Claim 2 (currently amended): A system according to
2 Claim 1, further comprising:
3 a document title capitalizer to provide one or
4 more of the words with an initial letter in uppercase
5 and each remaining letter in lowercase.

1 Claim 3 (original): A system according to Claim 1,
2 further comprising:
3 a sentence capitalizer to provide only an initial
4 such word with an initial letter in uppercase and each
5 remaining letter in lowercase.

1 Claim 4 (currently amended): A system according to
2 Claim 1, ~~further comprising:~~
3 ~~a word analyzer to skip at least one of each such~~
4 wherein the predefined type is one of (A) a word
5 comprising a number, ~~each such (B) a word~~ including no
6 vowels, and ~~each such (C) a word~~ not occurring at a
7 start of a phrase and constituting at least one of an
8 article, conjunction, preposition.

1 Claim 5 (original): A system according to Claim 1,
2 further comprising:
3 a lexicon comprising one or more reference words
4 with at least one reference word defining a form of
5 capitalization for the reference word;
6 a matcher to match the at least one such word
7 against the reference words, the evaluator skipping
8 each such word if a matching reference word is found.

1 Claim 6 (currently amended): A system according to
2 Claim 1, further comprising:
3 a proper noun capitalizer to provide each of the
4 individual letters in each such word comprising a noun
5 with no vowels in uppercase.

1 Claim 7 (currently amended): A system according to
2 Claim 1, ~~further comprising:~~
3 ~~a tokenizer to tokenize~~ wherein the preprocessor
4 tokenizes the excerpt into the one or more words and
5 one or more punctuation marks.

1 Claim 8 (currently amended): A method for providing
2 capitalization correction for unstructured excerpts,
3 comprising:
4 tokenizing an excerpt of unstructured content
5 into a set of words; and
6 analyzing the set of words for correct
7 capitalization, comprising:
8 evaluating individual characters constituting at
9 least one such word in the set of words; and
10 skipping the at least one such word if determined
11 to be of a predefined type such that the capitalizer
12 does not capitalize the at least one such word.

1 Claim 9 (original): A method according to Claim 8,
2 further comprising:
3 providing one or more of the words with an
4 initial letter in uppercase and each remaining letter
5 in lowercase.

1 Claim 10 (original): A method according to Claim 8,
2 further comprising:
3 providing only an initial such word with an
4 initial letter in uppercase and each remaining letter
5 in lowercase.

1 Claim 11 (currently amended): A method according to
2 Claim 8, ~~further comprising:~~
3 ~~skipping at least one of each such wherein the~~
4 predefined type is one of (A) a word comprising a

5 number, ~~each such~~ (B) a word including no vowels, and
6 ~~each such~~ (C) a word not occurring at a start of a
7 phrase and constituting at least one of an article,
8 conjunction, preposition.

1 Claim 12 (original): A method according to Claim 8,
2 further comprising:
3 maintaining a lexicon comprising one or more
4 reference words with at least one reference word
5 defining a form of capitalization for the reference
6 word;
7 matching the at least one such word against the
8 reference words; and
9 skipping each such word if a matching reference
10 word is found.

1 Claim 13 (currently amended): A method according to
2 Claim 8, further comprising:
3 providing each of the individual letters in each
4 such word comprising a noun with no vowels in
5 uppercase.

1 Claim 14 (currently amended): A method according to
2 Claim 8, ~~further comprising:~~ wherein the act of
3 tokenizing includes tokenizing the excerpt into the
4 one or more words and one or more punctuation marks.

Claim 15 (canceled)

1 Claim 16 (currently amended): An apparatus for
2 providing capitalization correction for unstructured
3 excerpts, comprising:
4 means for tokenizing an excerpt of unstructured
5 content into a set of words; and
6 means for analyzing the set of words for correct
7 capitalization, comprising:
8 means for evaluating individual characters
9 constituting at least one such word in the set of
10 words; and
11 means for skipping the at least one such word if
12 determined to be of a predefined type such that the
13 capitalizer does not capitalize the at least one such
14 word.

1 Claim 17 (currently amended): A system for building a
2 lexicon for use in capitalization correction for
3 unstructured excerpts, comprising:
4 a ripper assembling a list of word sets from
5 unstructured content, each word set comprising a word
6 and at least one variation on capitalization;
7 an aggregator aggregating each word set,
8 comprising:
9 an analyzer identifying at least one word set
10 comprising significant statistics; and
11 a non-standard capitalization selector selecting
12 at least ~~one~~ two such ~~variation~~ variations within the
13 identified word set having a non-standard
14 capitalization, and adding the at least ~~one~~ two such
15 ~~variation~~ variations to the lexicon.

1 Claim 18 (original): A system according to Claim 17,
2 further comprising:

3 a tokenizer tokenizing the excerpt into the one
4 or more words and one or more punctuation marks.

1 Claim 19 (original): A system according to Claim 18,
2 wherein hyphenated words are split into a plurality of
3 the words.

1 Claim 20 (original): A system according to Claim 17,
2 wherein at least one variation appearing at the start
3 of a sentence is skipped.

1 Claim 21 (original): A system according to Claim 20,
2 wherein the non-standard capitalization comprises the
3 at least one variation occurring in an excerpt having
4 fewer than half of individual letters provided in
5 uppercase.

1 Claim 22 (currently amended): A system according to
2 Claim 17, further comprising:
3 a normalizer normalizing a plurality of the words
4 extracted relative to a source of the unstructured
5 excerpt ~~structured Web content~~.

1 Claim 23 (currently amended): A system according to
2 Claim 17, wherein the set comprising significant
3 statistics comprises only non-standard capitalization
4 variations having at least four occurrences of at
5 least one such variation within a word set.

1 Claim 24 (original): A system according to Claim 17,
2 wherein the non-standard capitalization comprises the
3 at least one variation having any individual letter
4 other than the first individual letter provided in
5 uppercase.

1 Claim 25 (currently amended): A system according to
2 Claim 17, further comprising:
3 a standard capitalization selector selecting at
4 least ~~one~~ two such ~~variation~~ variations within the
5 identified word set having a standard capitalization,
6 and adding the at least ~~one~~ two such ~~variation~~
7 variations to the lexicon.

1 Claim 26 (currently amended): A system according to
2 Claim 17, further comprising:
3 a validator applying implicit rules for
4 capitalization, and skipping each of the at least ~~one~~
5 ~~variation~~ two variations subject to at least one such
6 implicit rule.

1 Claim 27 (currently amended): A system according to
2 Claim 26, wherein the implicit rules comprise skipping
3 each of the at least ~~one variation~~ two variations
4 based on position within a sentence or phrase.

1 Claim 28 (currently amended): A system according to
2 Claim 26, wherein the implicit rules comprise at least
3 one of (A) a number, (B) having no vowels, and (C)
4 constituting at least one of an article, conjunction
5 and preposition.

1 Claim 29 (currently amended): A system according to
2 Claim 26, wherein the implicit rules comprise
3 normalizing a number of occurrences for each of the at
4 least two variations ~~one variation~~ using at least one
5 of a normalizing function and relative to a source of
6 the each of the at least two variations ~~one variation~~.

1 Claim 30 (currently amended): A system according to
2 Claim 26, wherein the implicit rules comprise
3 accommodating multiple forms of capitalization for
4 each of the at least two variations ~~one variation~~ by
5 annotating each capitalization form with a frequency
6 count and skipping those of the each of the at least
7 two variations ~~one variation~~ occurring infrequently.

1 Claim 31 (original): A system according to Claim 17,
2 further comprising:
3 a hash table maintaining the lexicon.

1 Claim 32 (original): A system according to Claim 31,
2 further comprising:
3 at least one record specifying at least one such
4 word as a key into the hash table, and associating at
5 least one such variation within the word set as a
6 preferred capitalization.

1 Claim 33 (currently amended): A method for building a
2 lexicon for use in capitalization correction for
3 unstructured excerpts, comprising:
4 assembling a list of word sets from unstructured
5 content, each word set comprising a word and at least
6 one variation on capitalization;
7 aggregating each word set, comprising:
8 identifying at least one word set comprising
9 significant statistics;
10 selecting at least ~~one~~ two such ~~variation~~
11 variations within the identified word set having a
12 non-standard capitalization; and
13 adding the at least ~~one~~ two such ~~variation~~
14 variations to the lexicon.

1 Claim 34 (original): A method according to Claim 33,
2 further comprising:

3 tokenizing the excerpt into the one or more words
4 and one or more punctuation marks.

1 Claim 35 (original): A method according to Claim 34,
2 further comprising:
3 splitting hyphenated words into a plurality of
4 the words.

1 Claim 36 (original): A method according to Claim 33,
2 further comprising:
3 skipping at least one variation which may be at
4 the start of a sentence.

1 Claim 37 (original): A method according to Claim 36,
2 wherein the non-standard capitalization comprises the
3 at least one variation occurring in an excerpt having
4 fewer than half of individual letters provided in
5 uppercase.

1 Claim 38 (currently amended): A method according to
2 Claim 33, further comprising:
3 normalizing a plurality of the words extracted
4 relative to a source of the unstructured excerpt
5 ~~structured Web content~~.

1 Claim 39 (currently amended): A method according to
2 Claim 33, wherein the set comprising significant
3 statistics comprises only non-standard capitalization

4 variations having at least four occurrences of at
5 least one such variation within a word set.

1 Claim 40 (original): A method according to Claim 33,
2 wherein the non-standard capitalization comprises the
3 at least one variation having any individual letter
4 other than the first individual letter provided in
5 uppercase.

1 Claim 41 (currently amended): A method according to
2 Claim 33, further comprising:
3 selecting at least one such variation within the
4 identified word set having a standard capitalization,
5 and adding the at least ~~one~~ two such ~~variation~~
6 variations to the lexicon.

1 Claim 42 (original): A method according to Claim 33,
2 further comprising:
3 applying implicit rules for capitalization; and
4 skipping each at least one variation subject to
5 at least one such implicit rule.

1 Claim 43 (original): A method according to Claim 42,
2 wherein the implicit rules comprise skipping at least
3 one variation based on position within a sentence or
4 phrase.

1 Claim 44 (currently amended): A method according to
2 Claim 42, wherein the implicit rules comprise at least
3 one of (A) a number, (B) having no vowels, and (C)
4 constituting at least one of an article, conjunction
5 and preposition.

1 Claim 45 (currently amended): A method according to
2 Claim 42, wherein the implicit rules comprise
3 normalizing a number of occurrences for each of the at
4 least two variations ~~one variation~~ using at least one
5 of a normalizing function and relative to a source of
6 each of the at least two variations ~~one variation~~.

1 Claim 46 (currently amended): A method according to
2 Claim 42, wherein the implicit rules comprise
3 accommodating multiple forms of capitalization for
4 each at least one variation by annotating each
5 capitalization form with a frequency count and
6 skipping those of ~~the~~ each of the at least two
7 variations ~~one variation~~ occurring infrequently.

1 Claim 47 (original): A method according to Claim 33,
2 further comprising:
3 maintaining the lexicon structured as a hash
4 table.

1 Claim 48 (original): A method according to Claim 47,
2 further comprising:
3 specifying at least one such word as a key into
4 the hash table; and
5 associating at least one such variation within
6 the word set as a preferred capitalization.

Claim 49 (canceled)

1 Claim 50 (original): An apparatus for building a
2 lexicon for use in capitalization correction for
3 unstructured excerpts, comprising:
4 means for assembling a list of word sets from
5 unstructured content, each word set comprising a word
6 and at least one variation on capitalization;
7 means for aggregating each word set, comprising:
8 means for identifying each word set comprising
9 significant statistics;
10 means for selecting at least ~~one~~ two such
11 ~~variation~~ variations within the identified word set
12 having a non-standard capitalization; and
13 means for adding the at least ~~one~~ two such
14 ~~variation~~ variations to the lexicon.